Gene involved in epigenetic gene silencing

The present invention relates to DNA which encodes proteins that control gene silencing, and particularly the silencing of plant genes.

The loss of expression of previously active genes in plants, also referred to as gene silencing, is observed in response to developmental, environmental or unknown signals. It occurs at a frequency higher than that of mutations, yet it is markedly stable during somatic transmission. Gene silencing, initially perceived as an unwanted source of instability of transgene expression, is now regarded as a molecular tool to intentionally regulate gene expression.

It appears that chromosomal position or structure of the affected loci are factors determining the frequency and strength of silencing. Inactivation seems to preferentially affect genes present in multiple copies and is thought to be a consequence of sequence redundancy. Many examples of homology-dependent gene silencing have been reported. Closer analysis has allowed the classification of silencing events according to the relative position of the affected loci (*cis*, *trans*, allelic, ectopic), the origin of the affected genes (endogenous or transgenic), and the level of interaction (transcriptional or post-transcriptional). While post-transcriptional silencing seems to mainly involve the formation of aberrant RNA molecules and is occasionally, but not necessarily, accompanied by DNA methylation, silencing interfering with transcription initiation is more strictly correlated with hypermethylation of the DNA and possibly with alteration of chromatin structure at the silent loci. It is, however, not clear whether these molecular events are a prerequisite for gene silencing or a consequence of the silent state.

In the case of transcriptional silencing, the inactive state of silenced genes is stably transmitted through mitotic and meiotic divisions. As in other organisms, trans-acting modifier loci are assumed to be responsible for the stability of the inactive state of the silenced genes. Mutations in such loci resulting in mutated proteins are expected to result in reduced gene silencing and reactivation of previously silent loci by interfering with the maintenance of the silent state, or by a failure to recognize sequence redundancy. It has been reported that mutations in the DDM1 gene of Arabidopsis thaliana release transcriptional gene silencing and that this genes encodes a SWI2/SNF2-like protein involved in chromatin remodeling. However, mutation of the DDM1 gene causes severe pleiotropic effects. Therefore, to be able to modify such effects making use of gene technology, it is necessary to identify further specific modifier loci and characterize the corresponding wild-type and mutant proteins. It is the main objective of the present invention to provide DNA comprising an open reading frame encoding such a protein.

Trans-acting modifier loci according to the present invention can be identified by T-DNA insertion mutagenesis as described in Example 1 for an Arabidopsis line carrying a heritably inactivated, methylated hygromycin resistance gene. A mutation of a silencing modifier locus results in release of silencing of the hygromycin resistance gene and restores hygromycin resistance. Plants homozygous for the silent resistance gene are subjected to transformation with a selectable marker gene different from the hygromycin resistance gene, which is under the control of the T-DNA 1'-2' dual promoter. Transformants are selected and their progeny screened for hygromycin resistance. The mutant phenotype (hygromycin resistance) is screened for genetic co-segregation with a specific T-DNA insert. Cloning of the tagged gene using routine methods of recombinant DNA technology allows to characterize the mutant and wild-type DNA sequence of the silencing modifier locus as well as the encoded protein.

Within the context of the present invention reference to a gene is to be understood as reference to a DNA coding sequence associated with regulatory sequences, which allow transcription of the coding sequence into RNA such as mRNA, rRNA, tRNA, snRNA, sense RNA or antisense RNA. Examples of regulatory sequences are promoter sequences, 5' and 3' untranslated sequences, introns, and termination sequences.

A promoter is understood to be a DNA sequence initiating transcription of an associated DNA sequence, and may also include elements that act as regulators of gene expression such as activators, enhancers, or repressors.

Expression of a gene refers to its transcription into RNA or its transcription and subsequent translation into protein within a living cell. In the case of antisense constructs expression refers to the transcription of the antisense DNA only.

The term transformation of cells designates the introduction of nucleic acid into a host cell, particularly the stable integration of a DNA molecule into the genome of said cell.

Any part or piece of a specific nucleotide or amino acid sequence is referred to as a component sequence.

DNA according to the present invention comprises an open reading frame encoding a protein characterized by an amino acid sequence comprising a component sequence of at least 150 amino acid residues having 40% or more identity with SEQ ID NO: 3. In particular the protein encoded by the open reading frame can be described by the formula R_1-R_2-R_3, wherein

-- R_1, R_2 and R_3 constitute component sequences consisting of amino acid residues independently selected from the group of the amino acid residues Gly, Ala, Val, Leu, Ile, Phe, Pro, Ser, Thr, Cys, Met, Trp, Tyr, Asn, Gln, Asp, Glu, Lys, Arg, and His,

-- R_1 and R_3 consist independently of 0 to 3000 amino acid residues;

-- R_2 consists of at least 150 amino acid residues; and

-- R_2 is at least 40% identical to an aligned component sequence of SEQ ID NO: 3.

In most cases the total length of the protein will be in the range of 1000 to 3000 amino acid residues. In preferred embodiments of the invention the component sequence R_2 consists of at least 200 amino acid residues. Specific examples of the component sequence R_2 are component sequences of SEQ ID NO: 3 represented by the following range of amino acids:

1 - 416 (corresponding to exon 2);
418 - 583 (corresponding to exons 3 to 5);
584 - 890 (corresponding to exon 6);
892 - 1472 (corresponding to exons 7 to 9);
1007 - 1472 (corresponding to exon 9);
1473 - 1631 (corresponding to exons 10 to 12);
1632 - 1827 (corresponding to exons 13 to 15); and
1829 - 2001 (corresponding to exon 16).

In a preferred embodiment of the present invention at least one of the component sequences R_1 or R_3 comprises one or more additional component sequences with a length of at least 50 amino acids and at least 60% identical to an aligned component sequence of SEQ ID NO: 3. Specific examples of such additional component sequences are component sequences of SEQ ID NO: 3 represented by the following range of amino acids:

420 - 525 (corresponding to exons 3 and 4);
444 - 525 (corresponding to exon 4);
526 - 583 (corresponding to exon 5);
892 - 971 (corresponding to exon 7);
892 - 1006 (corresponding to exons 7 and 8);
1473 - 1524 (corresponding to exon 10);
1525 - 1576 (corresponding to exon 11);
1577 - 1631 (corresponding to exon 12);
1632 - 1690 (corresponding to exons 13);
1692 - 1757 (corresponding to exons 14); and
1758 - 1827 (corresponding to exons 15).

Particularly preferred embodiments of the DNA according to the present invention encode a protein having a component sequence defined by amino acids 478-490, 584-600, 617-630, 654-668, 676-690, 718-734, 776-788, 1222-1233, 1738-1749 or 1761-1770 of SEQ ID NO: 3. Preferably, the encoded protein comprises at least two, three or more different representatives of said component sequences. Specific examples of said embodiments encode a protein characterized by the amino acid sequence of SEQ ID NO: 3, an allelic amino acid sequence having amino acid residue K instead of M at position 705 of SEQ ID NO: 3, or an amino acid residue D instead of E at position 1219 of SEQ ID NO: 3.

Dynamic programming algorithms yield different kinds of alignments. In general there exist two approaches towards sequence alignment. Algorithms as proposed by Needleman & Wunsch and by Sellers align the entire length of two sequences providing a global alignment of the sequences. The Smith-Waterman algorithm on the other hand yields local alignments. A local alignment aligns the pair of regions within the sequences that are most similar given the choice of scoring matrix and gap penalties. This allows a database search to focus on the most highly conserved regions of the sequences. It also allows similar domains within sequences to be identified. To speed up alignments using the Smith-Waterman algorithm both BLAST (Basic Local Alignment Search Tool) and FASTA place additional restrictions on the alignments.

Within the context of the present invention alignments are conveniently performed using BLAST, a set of similarity search programs designed to explore all of the available sequence databases regardless of whether the query is protein or DNA. Version BLAST 2.0 (Gapped BLAST) of this search tool has been made publicly available on the internet (currently http://www.ncbi.nlm.nih.gov/BLAST/). It uses a heuristic algorithm which seeks local as opposed to global alignments and is therefore able to detect relationships among sequences which share only isolated regions. The scores assigned in a BLAST search have a well-defined statistical interpretation. Particularly useful within the scope of the present invention are the blastp program allowing for the introduction of gaps in the local sequence alignments and the PSI-BLAST program, both programs comparing an amino acid query sequence against a protein sequence database, as well as a blastp variant program allowing local alignment of two sequences only. Said programs are preferably run with optional parameters set to the default values.

Sequence alignments using BLAST can also take into account whether the substitution of one amino acid for another is likely to conserve the physical and chemical properties necessary to maintain the structure and function of the protein or is more likely to disrupt essential structural and functional features of a protein. Such sequence similarity is quantified in terms of a percentage of "positive" amino acids, as compared to the percentage of identical amino acids and can help assigning a protein to the correct protein family in border-line cases.

Sequence alignments using such computer programs reveal the presence of an ATP/GTP-binding motif A (amino acids 460 to 467 in SEQ ID NO:3), the consensus sequence of which is (Ala/Gly)XaaXaaXaaXaaGlyLys(Ser/Thr), wherein (Ala/Gly) indicates Ala or Gly, Xaa indicates any naturally occurring amino acid and (Ser/Thr) indicates Ser or Thr. Alignment additionally reveals a region (amino acid position 479 to 719 in SEQ ID: 3), similar to part of the ATPase/helicase domain of proteins in the SWI2/SNF2 family which are involved in chromatin remodeling but no significant overall sequence identity with known proteins.

Specific examples of DNA according to the present invention are described in SEQ ID NO: 1 and SEQ ID NO: 2 encoding an Arabidopsis protein described in SEQ ID NO: 3. Stretches of SEQ ID NO: 3 having 50 to 500 amino acids length can show between 20 and 50% sequence identity to stretches of known protein sequences after alignment. Overall alignments of SEQ ID NO: 3, however, result in sequence identities lower than 30%. Thus, the present invention defines a new protein family the members of which are characterized by an amino acid sequence comprising a component sequence of at least 150 amino acid residues having 40% or more identity with an aligned component sequence of SEQ ID NO: 3. Preferably the amino acid sequence identity is higher than 50% or even higher than 55%.

DNA encoding proteins belonging to the new protein family according to the present invention can be isolated from monocotyledonous and dicotyledonous plants. Preferred sources are corn, sugarbeet, sunflower, winter oilseed rape, soybean, cotton, wheat, rice, potato, broccoli, cauliflower, cabbage, cucumber, sweet corn, daikon, garden beans, lettuce, melon, pepper, squash, tomato, or watermelon. However, they can also be isolated from mammalian sources such as mouse or human tissues. The following general method, can be used, which the person skilled in the art knows to adapt to the specific task. A single stranded fragment of SEQ ID NO: 1 or SEQ ID NO: 2 consisting of at least 15, preferably 20 to 30 or even more than 100 consecutive nucleotides is used as a probe to screen a DNA library for clones hybridizing to said fragment. The factors to be observed for hybridization are described in Sambrook et al, Molecular cloning: A laboratory manual, Cold Spring Harbor Laboratory Press, chapters 9.47-9.57 and 11.45-11.49, 1989. Hybridizing clones are sequenced and DNA of clones comprising a complete coding region encoding a protein characterized by an amino acid sequence comprising a component sequence of at least 150 amino acid residues having 40% or more sequence identity to SEQ ID NO: 3 is purified. Said DNA can then be further processed by a number of routine recombinant DNA techniques such as restriction enzyme digestion, ligation, or polymerase chain reaction analysis.

The disclosure of SEQ ID NO: 1 and SEQ ID NO: 2 enables a person skilled in the art to design oligonucleotides for polymerase chain reactions which attempt to amplify DNA fragments from templates comprising a sequence of nucleotides characterized by any continuous sequence of 15 and preferably 20 to 30 or more basepairs in SEQ ID NO: 1 or SEQ ID NO: 2. Said nucleotides comprise a sequence of nucleotides which represents 15 and preferably 20 to 30 or more basepairs of SEQ ID NO: 1 or SEQ ID NO: 2. Polymerase chain reactions performed using at least one such oligonucleotide and their amplification products constitute another embodiment of the present invention.

EXAMPLES:

Example 1: *T-DNA Insertion*

Transgenic line A of *Arabidopsis thaliana* ecotype Zürich with a transcriptionally silenced locus containing multiple copies of a chimeric hygromycin phosphotransferase gene (*hpt*) has been described in Mittelsten Scheid et al, Mol Gen Genet 228: 104-112, 1991 and Mittelsten Scheid et al, Proc Natl Acad Sci USA 93: 7114-7119, 1996. A homozygous, diploid genotype of said line is subjected to *Agrobacterium* mediated gene transfer by *in planta* vacuum infiltration (Bechtold et al., C R Acad Sci Paris Life Science 316: 1194-1199, 1993) generating more than 4000 independent T-DNA transformants. The binary vector with T-DNA consisting of the coding region of the *bar* gene transcriptionally fused to the 1' promoter (p1'barbi), the *Agrobacterium* strain (C58ClRif^R) and the transformation protocol are described by Mengiste et al, Plant J 12: 945-948, 1997. Transformants (T1 plants) are selected by repeated spraying of germinated seedlings with Basta solution (150 mg/l) and grown to maturity.

Example 2: *Mutant Selection*

Selfed seeds (T2 families) are collected from individual transformants. Prior to screening for revertants of the silenced phenotype, seeds are dried for one week at room temperature and cold-treated at 4°C for a minimum of one week. Pooled aliquots of approximately 1000 seeds (consisting of 50 seeds from 20 T2 families) are surface-sterilized twice (with 5% sodium hypochlorite containing 0.1% Tween 80) for 7 min and washed with sterile double-distilled water. For selection, each aliquot is plated on 14-cm Petri dishes containing 75 ml germination medium (according to Masson et al, Plant J 2: 829-933, 1992) solidified with 0.8% agar and containing 10 mg/l hygromycin B (Calbiochem). To ensure equal distribution during sowing, seeds are mixed with 30 ml of the same medium containing 0.4% agar. As positive control two seeds from a hygromycin-resistant line are sown at marked locations on each plate. Plates are cold-treated at 4°C for 2 days and subsequently subjected to alternating periods of 16 hours light at 21°C and 8 hours darkness at 16°C. Hygromycin resistance is evaluated each day for 8-15 days after sowing.

Example 3: ***Molecular and Genetic Analysis of the Mutant***

Following identification of 11 hygromycin-resistant seedlings in one of the pools, the families forming this pool are re-screened individually. One family contains approximately 25% hygromycin-resistant seedlings. Six resistant plantlets of this family are transferred to larger vessels containing germination medium without hygromycin. After rosette formation and development of the root system, plants are transferred to soil for further growth and seed setting. Prior to potting, tissue explants are taken from each plant to generate callus cultures on RCA medium (Table 1) with or without 10 mg/l hygromycin B. Callus cultures are used as a source of material for DNA and RNA analyses and for a further confirmation of hygromycin resistance in this tissue.
Genomic DNA is isolated using a CTAB based method as described by Mittelsten Scheid et al, Mol Gen Genet 244: 325-330, 1994, and incubated with restriction enzymes *BamHI, HpaII, MspII, DraI, EcoRV, RcaI* or *HindIII*. Total RNA is obtained using a RNAeasy kit (Qiagen) according to the supplier's recommendation. Southern and northern blot analysis are performed under conditions described by Church and Gilbert, Proc Natl Acad Sci USA 81: 1991-1995, 1984, using DNA fragments labeled with ^{32}P by random prime labeling. The coding region of the *hpt* gene, or DNA consisting of the P35S promoter, *hpt* coding and terminator region, or the coding region of the *bar* gene together with the 1' promoter are used as probes.

Northern blot analysis of 4 hygromycin-resistant siblings shows restoration of transcription of the *hpt* gene. Southern blot analysis of said siblings indicates that there is no detectable rearrangement within the complex *hpt* insert. The *hpt* transgene complex in the mutant is still hypermethylated like in the original line A, as judged by Southern blot analysis with the methylation-sensitive restriction enzymes *HpaII* and *MspI*, and by genomic sequencing of the promoter region after treatment with bisulfate. There is also no influence of the mutation on the methylation of repetitive genomic DNA in contrast to that observed for the *som* mutations.

The hygromycin-resistant plants, as well as non-selected siblings from the same family are grown to set seeds, checked for Basta resistance in the next generation, and scored for the number and size of the T-DNA inserts by Southern analysis. The results demonstrate that the original T-DNA transformant must have contained 2 T-DNA insertions segregating independently in the siblings. One insert co-segregates with the hygromycin resistant mutant phenotype. A plant homozygous for this insert and lacking the other T-DNA insert, is used for cloning the corresponding T-DNA insertion site.

Histochemical GUS staining of crosses between plants with mutant phenotype and the transgenic plant line GUS-TS (obtainable from Dr. H. Vaucheret, INRA, Versailles Cedex, France) of Arabidopsis thaliana ecotype Colombia containing a transcriptionally silenced locus with multiple copies of a chimeric beta-glucuronidase (gus) gene reveals reactivation of the silent GUS gene in the F2 progeny which are homozygous for the mom allele.

Inbreeding of plants with the mom1 mutant phenotype does not result in any morphological abnormalities even in the 9th generation of inbreeding. This is in contrast to the *som* mutants.

Backcrossing of the mutant phenotype of mom1 with line A (see example 1) results in immediate resilencing of the reacivated hpt gene upon introduction of a wild-type MOM allele in F1 hybrids. This also is in contrast to the *som* mutants.

Table 1: Composition of RCA medium

RCA medium	
MS macro 10 x	100 ml
B5 micro 1000 x	1 ml
ferric citrate	5 ml
NT vitamins 100 x	10 ml
sucrose	10 g
MES	5 ml
agar	10 g
NAA	0.1 mg
BAP	1 mg
pH 5.8 (KOH)	
ad 1 l	

MS macro 10 x	
potassium nitrate	19 g
ammonium nitrate	16.5 g
calcium chloride (x 2 H_2O)	4.4 g
magnesium sulfate (x 7 H_2O)	3.7 g
potassium dihydrogen phosphate	1.7 g
ad 1 l	

B5 micro 1000 x	
magnesium sulfate (x H_2O)	1000 mg
boric acid	300 mg
zinc sulfate (x 7 H_2O)	200 mg
potassium iodide	75 mg
sodium molybdate (x 2 H_2O)	25 mg
copper sulfate (x 5 H_2O)	2.5 mg
cobalt chloride (x 6 H_2O)	2.5 mg
ad 100 ml	

ferric citrate	
ammonium iron citrate	10 g
ad 1 l	

NT vitamins 100 x	
myo-inositol	1000 mg
thiamine HCl	10 mg
ad 1 l	

MES	
MES	14 g
pH 6 (NaOH)	
ad 100 ml	

Example 4: ***Cloning of the "Silencing Gene"***

Genomic DNA from the plant containing only the T-DNA co-segregating with the hygromycin resistant mutant phenotype is isolated. The DNA is subjected to TAIL (thermal asymmetric interlaced) PCR according to Liu et al, Plant J 8: 457-463, 1995, using 3 specific, nested primers close to the right border of the T-DNA (5'-CAT CTA CGG CAA TGT ACC AGC-3' (SEQ ID NO: 4), 5'-GAT GGG AAT TGG CTG AGT GGC-3' (SEQ ID NO: 5), 5'-CAG TTC CAA ACG TAA AAC GGC-3' (SEQ ID NO: 6)) which are directed outwards, and one of several degenerate primers which might bind in flanking plant DNA. Two out of the following seven degenerate primers

AD1	5'-NTC GAS TWT SGW GTT-3'	(Liu et al supra; SEQ ID NO: 7)
AD2	5'-NGT CGA SWG ANA WGA A-3'	(Liu et al supra; SEQ ID NO: 8)
AD3	5'-WGT GNA GWA NCA NAG A-3'	(Liu et al supra; SEQ ID NO: 9)
AD4	5'-WGG WAN CWG AWA NGC A-3'	(SEQ ID NO: 10)
AD5	5'-WCG WWG AWC ANG NCG A-3'	(SEQ ID NO: 11)
AD6	5'-WGC NAG TNA GWA NAA G-3'	(SEQ ID NO: 12)
AD7	5'-AWG CAN GNC WGA NAT A-3'	(SEQ ID NO: 13)

actually result in amplification of specific fragments. The larger one obtained using AD7 is cloned and sequenced. It contains 50 bp of the T-DNA and 275 bp of flanking plant DNA. In Southern blot analysis it is shown that this PCR fragment contains the plant DNA flanking the T-DNA. The PCR fragment is used to screen a genomic library (Stratagene) of wild type *Arabidopsis thaliana* ecotype Columbia. Three genomic clones hybridizing to the PCR fragment are identified. The genomic clones are further mapped with restriction enzymes, hybridized to the PCR fragment and aligned to each other. In one of the genomic clones obtained (p4A-11), the sequence found to flank the T-DNA of the insertion mutation is located approximately in the middle of the genomic sequence. An approximately 800 bp EcoRI-SalI fragment of p4A-11 is used to obtain the overlapping genomic clone p5-6, and an approximately 700 bp EcoRI fragment of p5-6 is used to obtain genomic clone p30-1 overlapping with p5-6. An approximately 700 bp HindIII fragment of p30-1 is used to obtain the genomic clone p33-19 overlapping with p30-1. Said clones are sequenced to design primers for RT-PCR. The approximately 700 bp EcoRI fragment of p5-6 is further used for screening of a cDNA library of wild type *Arabidopsis thaliana* ecotype Zurich according to

Elledge et al, Proc Natl Acad Sci USA 88: 1731-1735, 1991). Nine cDNA clones are obtained and the longest clone p17-8 having a length of 2.6 kb is sequenced.

Example 5: ***Sequence Analysis and Alignments***

Taking into account the large size of the Arabidopsis silencing gene cloned above it cannot be entirely excluded that the authentic nucleotide and amino acid sequences of the gene and protein, respectively, might deviate from the sequences given in SEQ ID NO: 1, SEQ ID NO: 2, and SEQ ID NO: 3 at a few positions due to mutations arising from the cloning procedure or due to ambiguities in the sequencing reactions. Additionally, sequencing of DNA derived from a different ecotype can reveal allelic differences. Thus, the sequences of SEQ ID NO: 1, SEQ ID NO: 2, and SEQ ID NO: 3 represent the corresponding genes and proteins of *Arabidopsis thaliana* ecotype Zurich, whereas genomic sequences obtained from *Arabidopsis thaliana* ecotype Columbia reveal two mismatches at nucleotide positions 4338 (A instead of T) and 6721 (T instead of G) of SEQ ID NO: 1, which result in an amino acid residue K instead of M at position 705 of SEQ ID NO: 3 and an amino acid residue D instead of E at position 1219 of SEQ ID NO: 3.

The 2.6 kb cDNA clone is analyzed sequentially from both ends and is shown to contain one large ORF as well as a 3' untranslated sequence.

Analysis of the genomic clones reveals that clones p4A-11 and p5-6 contain sequences homologous to the cDNA sequence as well as 7 intron sequences. Comparing the genomic sequences with the DNA sequences flanking the T-DNA insert, it turns out that the T-DNA insertion causes a deletion of about 2 kb of genomic DNA. The 5' end of the deletion is located in an intron (intron 12) and the 3' end of the deletion is located downstream of the 3' end of the cDNA. The sequence of 5' end of the cDNA clone terminates in the middle of the sequence of the genomic clone p5-6. Three independent nested RT-PCR reactions are performed to obtain additional cDNA sequences further upstream. The sequences of the primers used for these RT-PCRs are as follows:

RT1-1	5'-CTGTACATACTGAGTACAATCGGA-3'	(SEQ ID NO: 14)
RT1-2	5'-GCTTCAATTCCTGCCTCAGTTGAAC-3'	(SEQ ID NO: 15)
RT1-3	5'-CTCTACGTGCTTAACATCATGCGA-3'	(SEQ ID NO: 16)
RT1-4	5'-CCAGCTTCTGCTACTAGAAAGTCAG-3'	(SEQ ID NO: 17)

RT2/3-1	5'-CTGGAGTTGCATGAAATCCTGGATG-3'	(SEQ ID NO: 18)
RT2/3-2	5'-GCTCTTTGTAAGCTGTTCACGAGAC-3'	(SEQ ID NO: 19)
RT2-3	5'-TCGCATGATGTTAAGCACGTAGAG-3'	(SEQ ID NO: 20)
RT2-4	5'-GAGTACTGGTCCGTGAACAGGTAAT-3'	(SEQ ID NO: 21)
RT3-3	5'-ATGCTTGCACAAGCATGGTCGGAAA-3'	(SEQ ID NO: 22)
RT3-4	5'-TGCAACATCGTGCATTTGCTCCAGA-3'	(SEQ ID NO: 23)
RT4-1	5'-CACAAGCATGAGTTTTTCCTTCCGG-3'	(SEQ ID NO: 24)
RT4-2	5'-CTGACTTTCTAGTAGCAGAAGCTGG-3'	(SEQ ID NO: 25)

Sequences of several parts of the genomic clones are found to be deposited in the *Arabidopsis* database (accession numbers B67281, B62563, B20434, B20425, B21274, B08967, B11993, B20116, B12496 and B10852 as end sequences of BAC, and Z18494 and AA597930 as partial cDNA sequences, on 13 Apr 1999). A comparison of the encoded protein sequence with the Swiss Protein Database reveals partial similarity with ATPase/helicase proteins of the SWI2/SNF2 family (amino acid position 479 to 719 in SEQ ID NO: 3). The encoded protein consists of 2001 amino acids and is calculated to have a molecular weight of 219 kD and a pI of 5.1. An ATP/GTP-binding motif (amino acid position 460 to 467 in SEQ ID NO: 3) and three nuclear localization motifs (amino acid positions 362 to 367, 832 to 838 and 858 to 862 in SEQ ID NO: 3) are found in the encoded protein. Subcellular immunodetection of HA-tagged MOM protein confirms its nuclear localization. Similarity to the actin binding domain of chicken tensin (amino acid position 1899 to 1941 in SEQ ID NO: 3) and a predicted membrane spanning domain (amino acid position 995 to 1015 in SEQ ID NO: 3) are also detected. Additionally, the encoded protein contains three types of repetitive regions or internal repeats essentially defined by amino acid positions 177 to 350, 1462 to 1672 and 1848 to 1894 OF SEQ ID NO: 3.

Example 6: ***Homologous genes in other species***

A putative proline/hydroxyproline-rich glycoprotein of *Arabidopsis thaliana* showing partial similarity to the MOM protein is disclosed as GenBank accession nubmer AAD29829). The similarity is 34-47% depending on the region and is only seen in the second half of the MOM protein (i.e. amino acids 1368 to 1944).

The MOM cDNA clone is used to probe genomic DNA from turnip, tomato, tobacco, maize, mouse, fruit fly and man for the presence of homologous genes by Southern blot analysis. Hybridization under conditions of low stringency is found in all cases. Cross-hybridizing clones from libraries can be identified and sequenced.

A genomic library of the *Brassica oleracea* var. acephala (obtained from Dr. Mark Cock, INRA, CNRS, Lyon, France) are screened with the *MOM* cDNA under stringent conditions. Two positive clones are obtained, subcloned, and partially sequenced. Partial sequences of clone 1 show similarity to different regions in the *MOM* gene (80-86% at DNA level and 62-80% at amino acid level) which encode the N-terminal, ATPase, and C-terminal parts of the MOM protein. All three putative nuclear localization sequences of the MOM protein are fully conserved in clone 1. Partial sequences of clone 2 also show similarity regions in the *MOM* gene (64-76% at DNA level and 55-64% at amino acid level) which encode the ATPase, putative transmembrane, and C-terminal parts of the MOM protein. The sequences of clones 1 and 2 are not identical, suggesting the presence of, at least, two homologous genes in *Brassica oleracea.* Examples of partial sequences obtained from clone 1 and 2 are given in SEQ ID Nos: 26-33.

Additionally a genomic library of *Brassica rapa* (obtained from Dr. Kinya Toriyama, Tohoku University, Sendai, Japan) is screened with the *MOM* cDNA under stringent conditions. Positive signals hybridizing to both a 5' and a 3' part of the *MOM* cDNA are obtained.

Furthermore, a genomic library of *Petunia hybrida* (obtained from Dr. Jan Kooter, Vrije Universiteit, Amsterdam, The Netherlands) is screened with *MOM* cDNA under less stringent conditions. Positive signals hybridizing to both the 5' and 3' part of the *MOM* cDNA are obtained.

Example 7: ***Manipulating marker gene expression by antisense constructs***

The 2.6 kb cDNA fragment and a 1.8 kb RT-PCR fragment amplified by a nested RT-PCR using primers RT1-1 and RT1-2 for the first PCR and primers RT1-3 and RT1-4 for the second PCR, are each inversely cloned into the multiple cloning site of the binary vector pbarbi53 to generate antisense RNA. pbarbi53 is a modified vector of p1'barbi and carries an expression cassette consisting of the 35S promoter of cauliflower mosaic virus, a multiple cloning site containing Xho I, SnaBI, Hpa I and Cla I restriction sites and the 35S terminator of cauliflower mosaic virus at the HindIII site of p1'barbi. The resulting recombinant plasmids are introduced into Agrobacterium as described in Example 1. The transgenic plant line GUS-TS (obtainable from Dr. H. Vaucheret, INRA, Versailles Cedex, France) of Arabidopsis thaliana ecotype Colombia containing a transcriptionally silenced locus with multiple copies of a chimeric beta-glucuronidase (gus) gene, is transformed with the recombinant plasmids as described in Example 1 and transformants are selected as described by Mengiste et al, Plant J 12: 945-948, 1997. pbarbi53 vector DNA is used in control transformations. The transformants are examined for reactivation of the gus gene by histochemical staining. A cotyledon leaf is soaked in gus staining solution (100 mM sodium phosphate buffer (pH 7.0), 0.05% 5-bromo-4-chloro-3-indolyl-beta-D-glucuronidase, 0.1% sodium azide) under vacuum for 10 min and then incubated at 37ºC overnight. While strong gus activity is observed in the plants transformed with the recombinant plasmid carrying the 2.6 kb cDNA, plants transformed with the recombinant plasmid carrying the 1.8 kb RT-PCR fragment or pbarbi53 do not show any gus activity above background. Therefore, expression of the antisense RNA of the 2.6 kb cDNA mimicks the mutant phenotype and confirms that sequences shown in SEQ ID NO: 1, SEQ ID NO: 2 and SEQ ID NO: 3 represent the genetic information for a component of the transcriptional gene silencing system.